Summary Tapes and Basic Summary Tabulations (BSTs)

•••

DLI Ontario Training, December 4-6, 2017

Tomasz Mrozewski, Laurentian University

Goals of this presentation

- Background info on BSTs
- Determining contents and accessing data
- Walkthrough of operationalizing research questions with BSTs
- Highlight some tips, tricks and peculiarities

Participant outcomes

- Learn to access data and documentation
- Learn when to use BSTs vs. print resources
- Gain transferrable knowledge about operationalizing research questions using BSTs

What are BSTs?

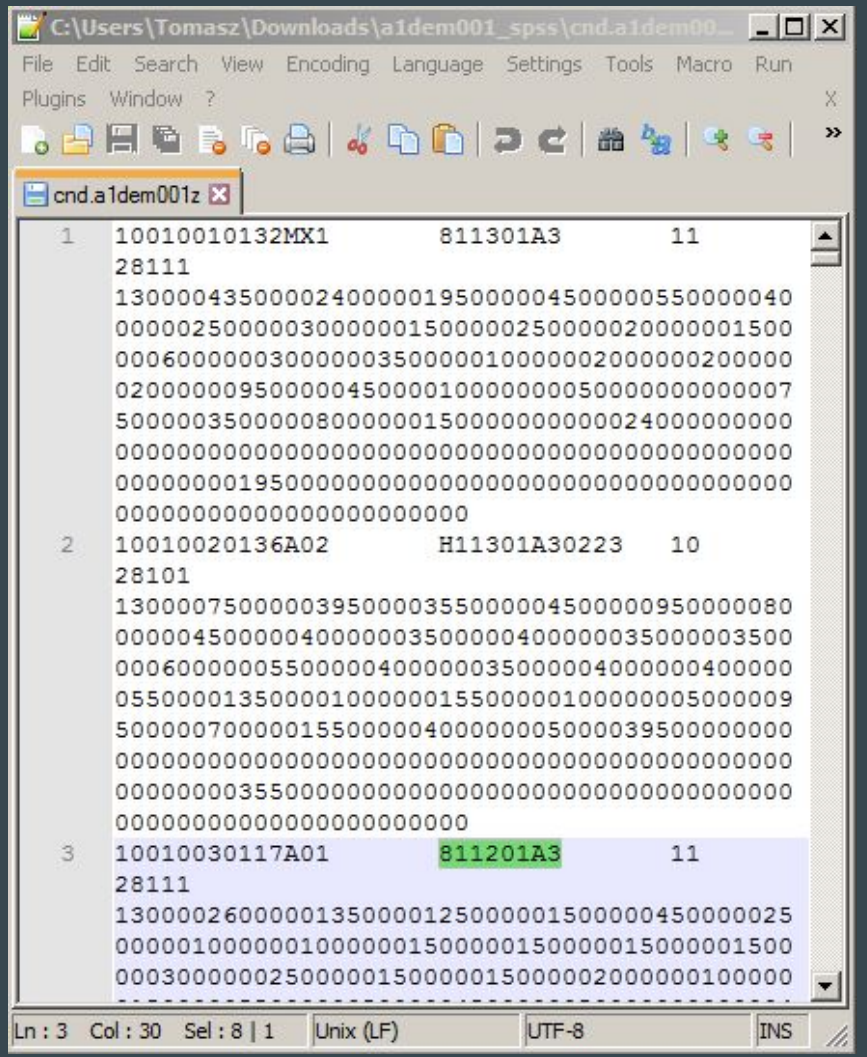
- Data tape files produced from 1961-1996; digitized by 2001.
- Becomes standard data product for Census: “this series provides cross-tabulations of two or more variables at fine levels of geographic detail.” (1996 Census Catalogue)
- Account of development given in *Administrative report of the 1961 Census*, “Chapter V: The tabulation programme”
- Also called “User Summary Tapes” in some years (1976, 1981)
- Replaced by Topic-Based Tabulations in 2001 (2001 Census Catalogue)

Structure of presentation

1. Background info
2. Accessing data and documentation
3. Walkthroughs of 3 research questions

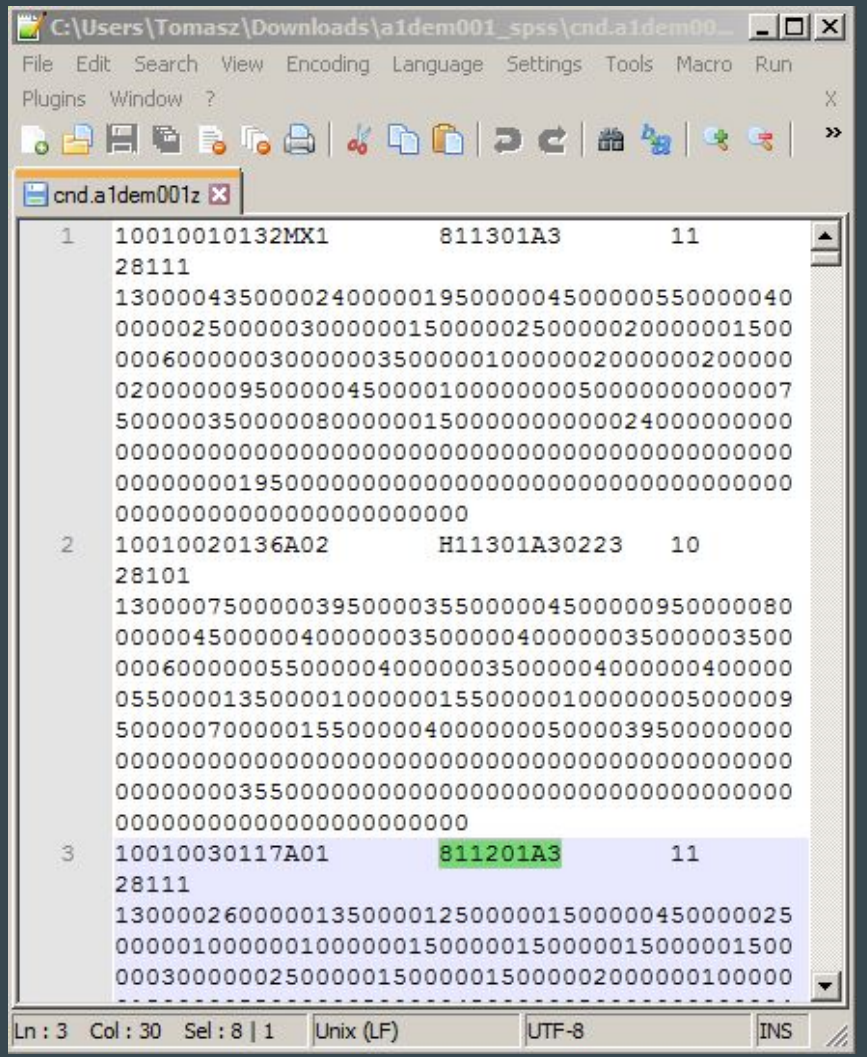
What do BSTs look like?

- Structured text files
- Filenames taken from tape names
- Readable with text editors like Notepad, Notepad++ (pictured), or EditPad
- Require syntax to make tabular
- Multiple tables in each tape/file
    - Values for all variables in all tables for same unit of observation in same row



What do BSTs look like? (cont'd)

- Each record/row *usually* a geographic unit
  - BSTs available at fixed levels:
    - Enumeration Area (EA)
    - Census Tract (CT)
    - Municipality 5,000+ (later, CSD)
    - CMA/CA
- Sometimes multiple record per geographic unit
  - EG, 602 MM-Z50 "Five-year age groups" (1966) has "maximum of sixteen records per E.A.: One record for each possible Combination of sex, residence, and marital status. (2 x 2 x 4). Records with a zero population In the particular... combination are dropped." (File descriptions of all Dominion Bureau of Statistics 1961 and 1966 Census summary tapes)


C:\Users\Tomasz\Downloads\a1dem001_spss\cnd.a1dem00...
File Edit Search View Encoding Language Settings Tools Macro Run
Plugins Window ?
cnd.a1dem001z
1 10010010132MX1 811301A3 11
28111
1300004350000240000019500000450000055000004000000250000030000001500000250000020000001500000600000030000003500000100000020000002000000200000095000004500001000000005000000000007500000350000080000001500000000002400000000000

How to find out what's in the BSTs

Apart from documentation:

- Census reports and catalogues
- UofT MDL site

Census reports and catalogues

Inconsistent coverage of BST descriptions

- 1981 - 1996: Census catalogues
- 1971 - 1976: no descriptions outside documentation
- 1961 - 1966: Administrative reports of Censuses

Census catalogues and administrative reports available in Internet Archive StatCan collection

How to access BST data

Two best sources of data:

- UofT Map and Data Library
- DLI EFT

Alternative, partial sources:

- CHASS (back to 1981)
- <odesi> (back to 1991; maybe including 1981 also)

UofT MDL

https://mdl.library.utoronto.ca/collections/numeric-data/census-canada

Advantages

- HTML-based directory of BSTs 1961-1976
- Syntax files and output (SPSS and SAS) often available, plus raw data, most documentation and some extras

Issues

- Migration to new site not complete, some broken links
- Some Official Lists (geography reference missing)
- Minor errors in some old syntax files

UofT MDL (cont'd)

Example: documentation available for 1966 BST

- Description of EA Summary Tape Files, 1966 (Statistics Canada)
- Census of Canada, 1961 - 1966: Information Guide for User Summary Tapes (Statistics Canada)
- File Descriptions of all Dominion Bureau of Statistics 1961 and 1966 Census Summary Tapes (Created by University of Alberta)
- Geographic reference files:
    - Official grouping of enumeration areas into townships, municipalities, parishes, cities, towns, villages, etc. (Atlantic provinces, Quebec, Ontario, Western provinces)
    - Official grouping of enumeration areas (Census Tracts, Cities from 10,000 and over (excluding census metropolitan areas) with an urban fringe)

UofT MDL (cont'd)

Example: BST data download from 1961 Census

- Codebook (ASCII)
- Record layout (ASCII)
- Tabular data file (SAV)
- SPSS syntax file (SPS)
- Data file (ASCII)

Remember: ASCII files can be opened with text editors

DLI EFT

Accessed via FTP: MAD_DLI_IDD_DAM > Root > census_pop_recens

- Organized by census year
- 1961: folder called “Cross-tabs”; 1966-1966: data in “BST” folder; documentation in “Doc” folder

Advantages:

- All data in raw, ASCII format - no tabular formats or syntax files
- Lots of original documentation, including Official Lists

Issues:

- No handy directory outside of documentation
- No tabular data or syntax

Geographic documentation

Available geographic reference products vary by year and may include:

- Reference maps
- Geography tape files and Geographic Attribute Files (GAF) - point files of EA locations
- Record layout and Codebook files (often only province and country/district/CD)
- Official lists

Note: Official lists the only way to definitively tie geographies to named locations

Operationalizing research questions

Walkthrough questions derived from a research project “to analyze demographic trends and the structure of the labour force in regions of Northern Ontario” by Economics profession David Leadbeater.

- Structure of labour force by CD, 1961-1996
- Class of worker by CD, 1961-1981
- Income by pre-amalgamation municipality, Timmins, 1961-1981

Walkthrough: Composition of labour force

Composition of labour force: size, participation rate, employment and unemployment rates, by sex and by age groups (15-64 and 65+), by CD

Walkthrough: Composition of labour force (cont'd)

What can we get from print materials? (1976)

- Labour force activity by marital status, age and sex : Canada, provinces, census metropolitan areas (no CD level data)
- Labour force activity by sex (no age data)
- Labour force participation rates by age and sex : Canada, provinces, census divisions (no labour force *activity* data)
- Labour force participation rates by age and sex, 1971 and 1976 : census divisions, municipalities of 5,000 population and over (ditto)

Walkthrough: Composition of labour force (cont'd)

Census Year	15 - 64 OR 65+ Years											
	Total			Participation rate			Employed			Unemployed		
	T	M	F	T	M	F	T	M	F	T	M	F
2011	y	y	y	y	y	y	y	y	y	y	y	y
2006	y	y	y	y	y	y	y	y	y	y	y	y
2001	y	y	y	y	y	y	y	y	y	y	y	y
1996	y	y	y	y	y	y	y	y	y	y	y	y
1991	C	C	C	C	C	C	C	C	C	C	C	C
1986	D	D	D	D	D	D	D	D	D	D	D	D
1981	y	y	y	y	y	y	y	y	y	y	y	y
1976	B	B	B	B	B	B	E	E	E	E	E	E
1971	B	B	B	B	B	B	-	-	-	-	-	-
1966	-	-	-	-	-	-	-	-	-	-	-	-
1961	-	-	-	-	-	-	-	-	-	-	-	-

Legend:	
A	Labour force activity by sex, table 2
B	Labour force participation rates by age and sex, 1971 and 1976, table 1
C	Table N9101 (Basic Summary Tabulation, 1991 census)
D	Table lf86b01
E	Table sdecob40
F	A2ECN001 table 10
G	Employment Status by sex, table 6
H	Labour force : earnings and employment of wage-earners, table 2

Walkthrough: Composition of labour force (cont'd)

1976: which data table do we need?

Via MDL 1976 Census page, under "Census Division/Subdivision Level – Long Form"

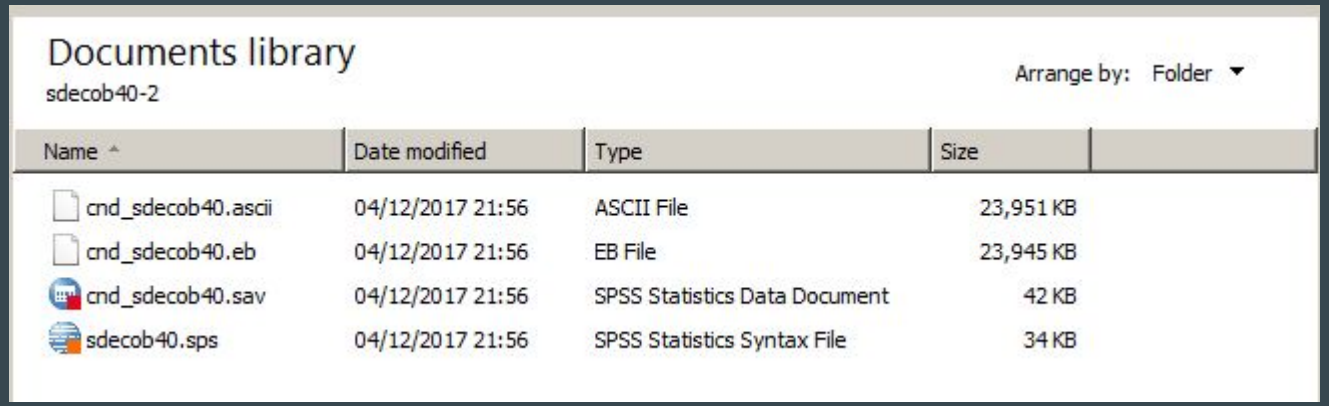
Economic file 2 (SDECOB20)

- Geographic levels: Census Division (CD), Census Subdivision (CSD)
- File contents:
  - Table SDECOB21: Population 15 years and over by labour force activity (7) by age (6) by level of schooling (5) by sex (3)
- Reference material:
  - Geographic identifiers
  - Record layout
  - Codebook
- Data: ASCII with SPSS/SAS control cards

Walkthrough: Composition of labour force (cont'd)

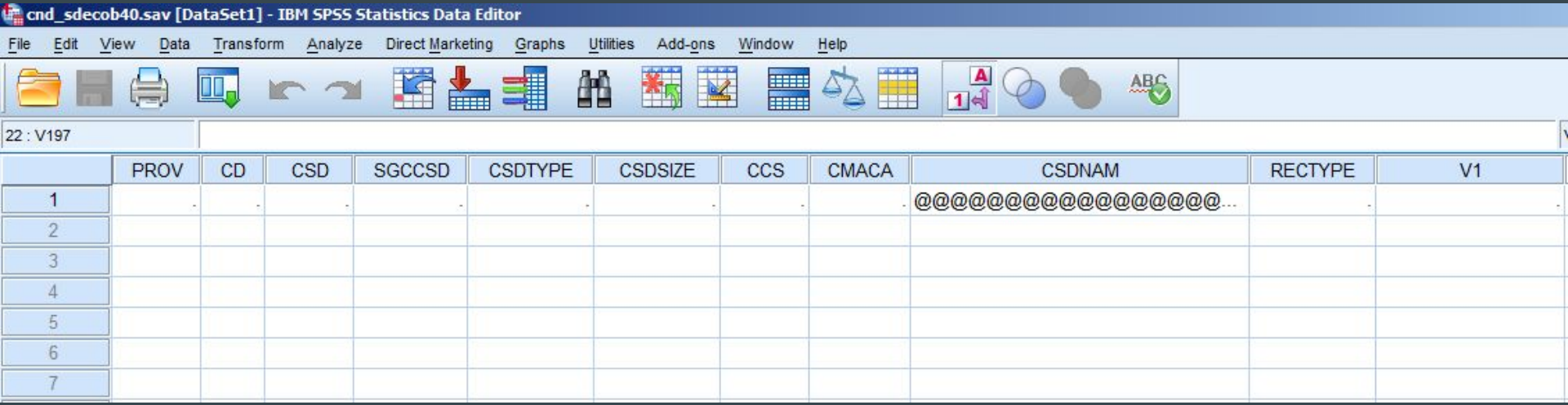
Folder contents:

- .eb and .ascii files are actually the same
- Codebook and layout file were available through download page (previous slide)
- Documentation not bundled for this BST



Walkthrough: Composition of labour force (cont'd)

Problem!

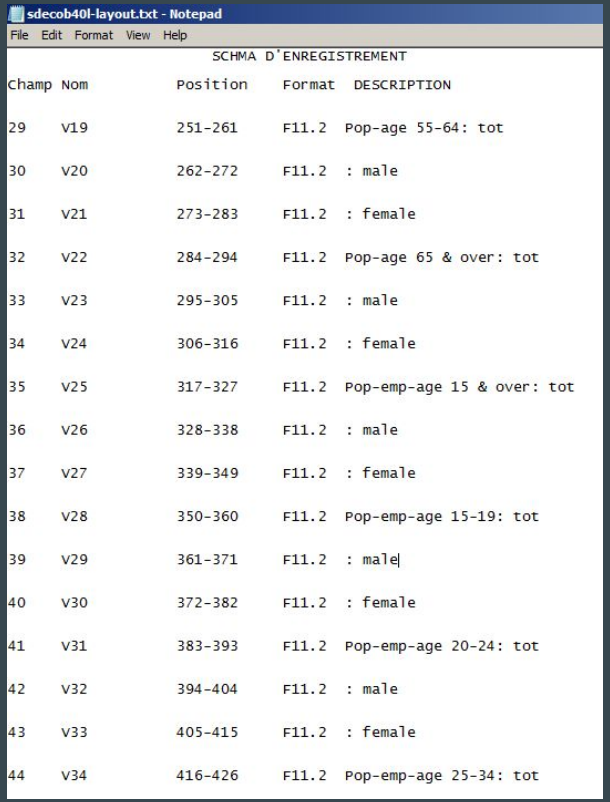


Fortunately, it was easily fixed by editing the syntax file and re-running it on the raw data.

Walkthrough: Composition of labour force (cont'd)

Data is available in semi-long format - see screenshot of layout file, right.

Fortunate: we have total labour force and labour force 65+ already given for all activity variables.



sdecob40l-layout.txt - Notepad

File Edit Format View Help

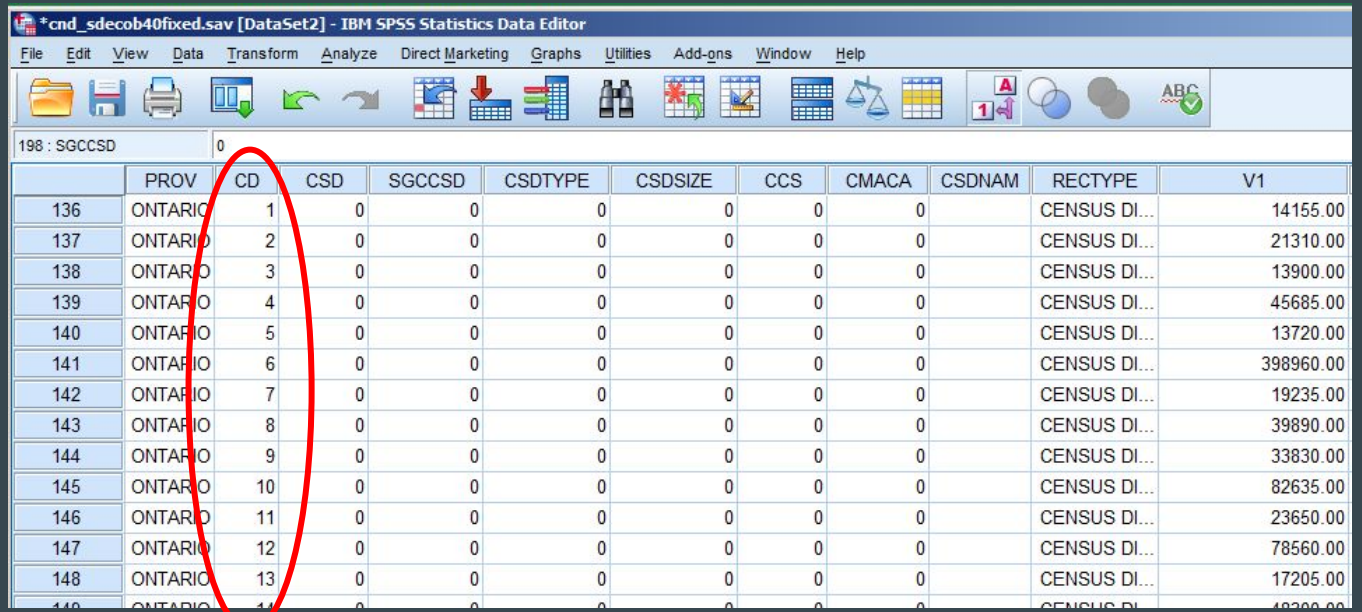
SCHMA D'ENREGISTREMENT

Champ	Nom	Position	Format	DESCRIPTION
29	V19	251-261	F11.2	Pop-age 55-64: tot
30	V20	262-272	F11.2	: male
31	V21	273-283	F11.2	: female
32	V22	284-294	F11.2	Pop-age 65 & over: tot
33	V23	295-305	F11.2	: male
34	V24	306-316	F11.2	: female
35	V25	317-327	F11.2	Pop-emp-age 15 & over: tot
36	V26	328-338	F11.2	: male
37	V27	339-349	F11.2	: female
38	V28	350-360	F11.2	Pop-emp-age 15-19: tot
39	V29	361-371	F11.2	: male
40	V30	372-382	F11.2	: female
41	V31	383-393	F11.2	Pop-emp-age 20-24: tot
42	V32	394-404	F11.2	: male
43	V33	405-415	F11.2	: female
44	V34	416-426	F11.2	Pop-emp-age 25-34: tot

Walkthrough: Composition of labour force (cont'd)

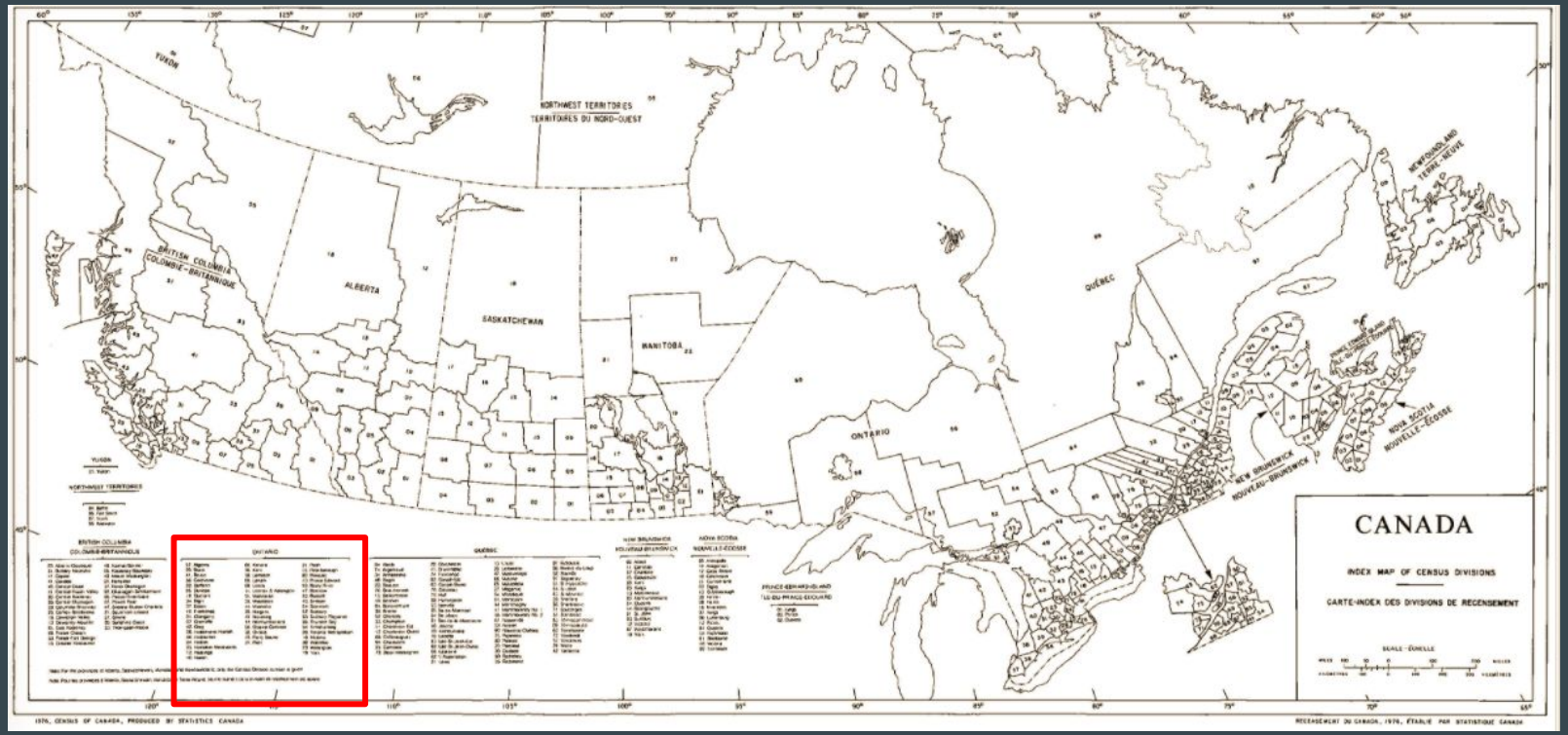
Next problem: no CD codes in syntax, output, codebook, or layout file!

Can we assume current SGC codes were applicable in 1976?



	PROV	CD	CSD	SGCCSD	CSDTYPE	CSDSIZE	CCS	CMACA	CSDNAM	RECTYPE	V1
136	ONTARIO	1	0	0	0	0	0	0		CENSUS DI...	14155.00
137	ONTARIO	2	0	0	0	0	0	0		CENSUS DI...	21310.00
138	ONTARIO	3	0	0	0	0	0	0		CENSUS DI...	13900.00
139	ONTARIO	4	0	0	0	0	0	0		CENSUS DI...	45685.00
140	ONTARIO	5	0	0	0	0	0	0		CENSUS DI...	13720.00
141	ONTARIO	6	0	0	0	0	0	0		CENSUS DI...	398960.00
142	ONTARIO	7	0	0	0	0	0	0		CENSUS DI...	19235.00
143	ONTARIO	8	0	0	0	0	0	0		CENSUS DI...	39890.00
144	ONTARIO	9	0	0	0	0	0	0		CENSUS DI...	33830.00
145	ONTARIO	10	0	0	0	0	0	0		CENSUS DI...	82635.00
146	ONTARIO	11	0	0	0	0	0	0		CENSUS DI...	23650.00
147	ONTARIO	12	0	0	0	0	0	0		CENSUS DI...	78560.00
148	ONTARIO	13	0	0	0	0	0	0		CENSUS DI...	17205.00

Walkthrough: Composition of labour force (cont'd)



Source: [Census divisions and subdivisions, Ontario]

Walkthrough: Class of worker

Class of worker data for Ontario CDs, 1961-1981.

Class of worker: definition varies per year but often includes wage earners, self-employed, and unpaid family workers

Walkthrough: Class of worker (cont'd)

What can we get from print materials? (1971)

- Class of worker and 1970 wage distribution (province and CMA)
- Employment income by sex, class of worker, age and weeks worked, full- and part-time
- Industries by sex, showing age, class of worker and number of non-Canadian born, for census metropolitan areas (Calgary-Quebec) and (Regina-Winnipeg)
- Industry divisions and major groups by sex, showing age, class of worker and number of married females, for census metropolitan areas (place of work)
- Occupations by sex, showing age, marital status and class of worker, for Quebec and Ontario

Walkthrough: Class of worker (cont'd)

Via MDL 1971 Census page, under "Enumeration area, long form."

Note: many more tables on class of worker available at CT and CMA level, but none others that can be rolled up to CD.

Economic file 1 (A2ECN001)

- Geographic level: Enumeration Area (EA)
- File contents:
  - Table 1: male population aged 15 and over by income (2)
  - Table 2: males aged 15 and over who worked in 1970. by employment income (2)
  - Table 3: male wage earners in labour force by wage and salary income (2)
  - Table 4: population aged 15 and over by marital status (4) by sex (2)
  - Table 5: labour force by marital status (4) by sex (2)
  - Table 6: population aged 15 and over by age (8) by sex (2)
  - Table 7: labour force by age (8) by sex (2)
  - Table 8: population aged 15 and over by level of schooling (81 by sex (2)
  - Table 9: labour force by level of schooling (8) by sex (2)
  - Table 10: population aged 15 and over by labour force status (12) by sex (2)
  - Table 11: experienced labour force by class of worker (3) by sex (2)
  - Table 12: experienced labour force by occupation (15) by sex (2)
  - Table 13: experienced labour force by industry (12) by sex (2)
- Documentation:
  - Record layout (ASCII)
  - Codebook (ASCII)
  - Complete documentation (pdf)
- Data
  - SPSS (contents: data file, syntax file .sps, system file .sav, all relevant documentation)
  - SAS (contents: data file, syntax file .sas, all relevant documentation)

Walkthrough: Class of worker (cont'd)

Variables available at EA level - from documentation PDF on download page (see previous slide)

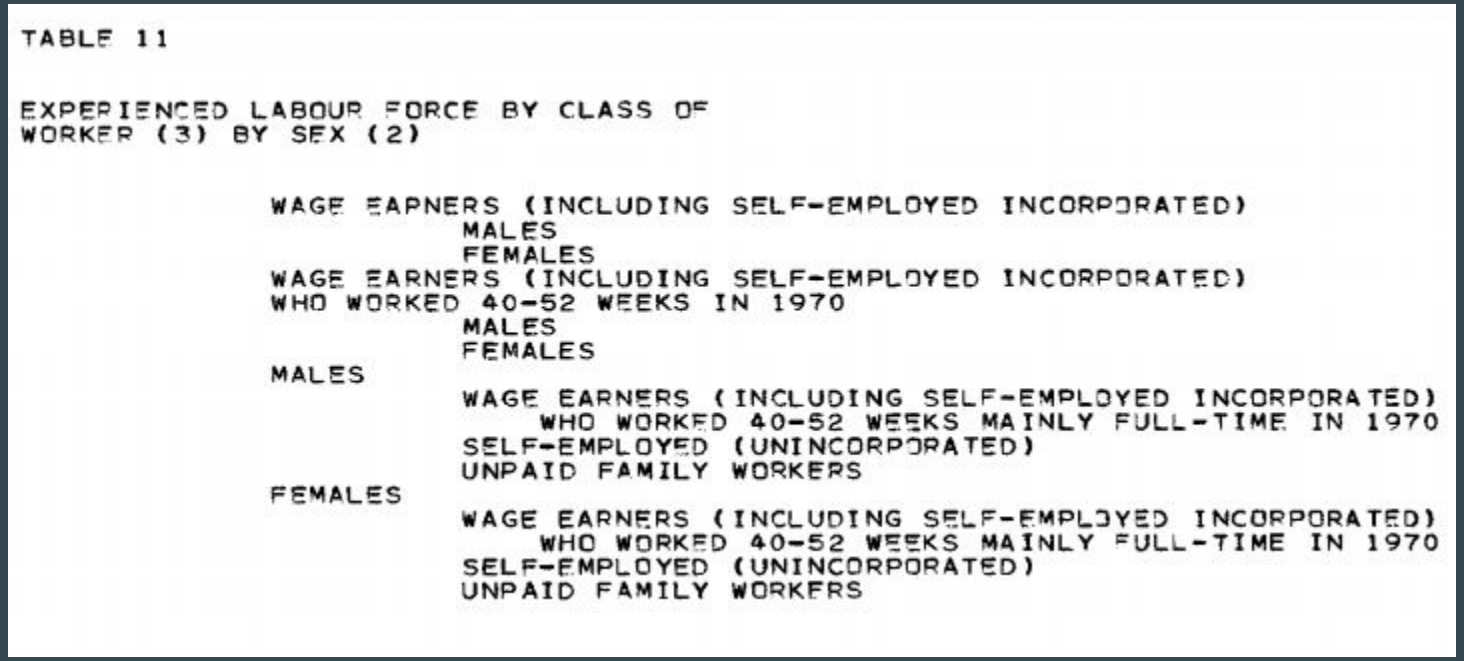


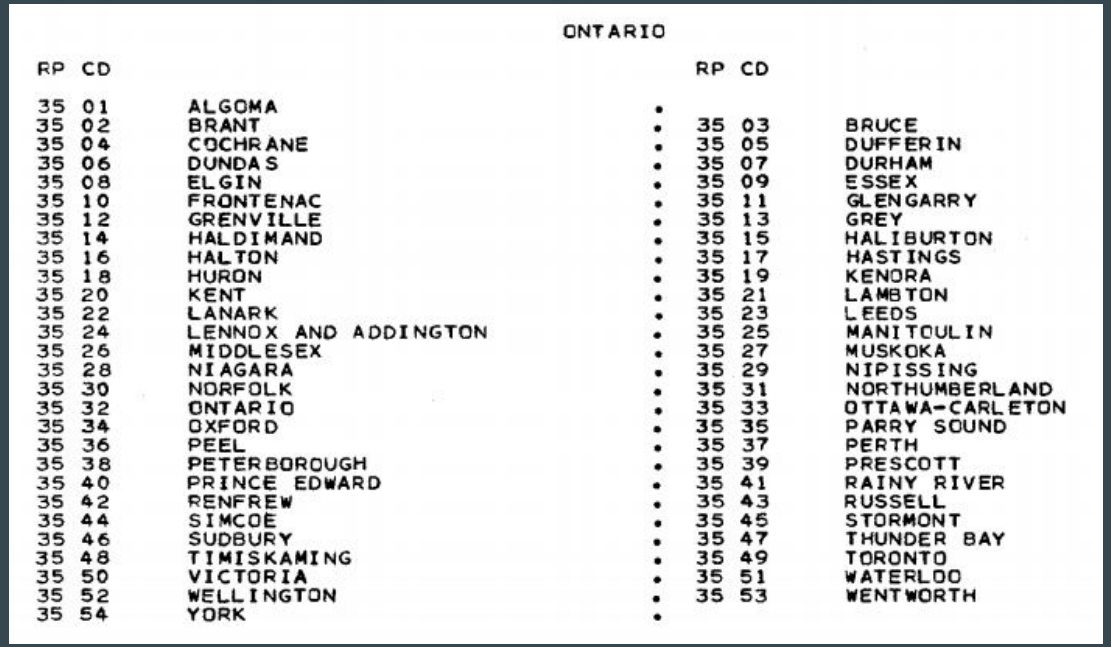
TABLE 11

EXPERIENCED LABOUR FORCE BY CLASS OF WORKER (3) BY SEX (2)

WAGE EARNERS (INCLUDING SELF-EMPLOYED INCORPORATED)
MALES
FEMALES
WAGE EARNERS (INCLUDING SELF-EMPLOYED INCORPORATED) WHO WORKED 40-52 WEEKS IN 1970
MALES
FEMALES
MALES
WAGE EARNERS (INCLUDING SELF-EMPLOYED INCORPORATED) WHO WORKED 40-52 WEEKS MAINLY FULL-TIME IN 1970
SELF-EMPLOYED (UNINCORPORATED)
UNPAID FAMILY WORKERS
FEMALES
WAGE EARNERS (INCLUDING SELF-EMPLOYED INCORPORATED) WHO WORKED 40-52 WEEKS MAINLY FULL-TIME IN 1970
SELF-EMPLOYED (UNINCORPORATED)
UNPAID FAMILY WORKERS

Walkthrough: Class of worker (cont'd)

Identifying the CDs, also from documentation PDF ("List of Counties and Census Subdivisions) - note that the CD numbers are different than 1976 CDs in previous walkthrough.



ONTARIO

RP	CD		RP	CD	
35	01	ALGOMA			
35	02	BRANT	35	03	BRUCE
35	04	COCHRANE	35	05	DUFFERIN
35	06	DUNDAS	35	07	DURHAM
35	08	ELGIN	35	09	ESSEX
35	10	FRONTENAC	35	11	GLENGARRY
35	12	GRENVILLE	35	13	GREY
35	14	HALDIMAND	35	15	HALIBURTON
35	16	HALTON	35	17	HASTINGS
35	18	HURON	35	19	KENORA
35	20	KENT	35	21	LAMBTON
35	22	LANARK	35	23	LEEDS
35	24	LENNOX AND ADDINGTON	35	25	MANITOULIN
35	26	MIDDLESEX	35	27	MUSKOKA
35	28	NIAGARA	35	29	NIPISSING
35	30	NORFOLK	35	31	NORTHUMBERLAND
35	32	ONTARIO	35	33	OTTAWA-CARLETON
35	34	OXFORD	35	35	PARRY SOUND
35	36	PEEL	35	37	PERTH
35	38	PETERBOROUGH	35	39	PRESCOTT
35	40	PRINCE EDWARD	35	41	RAINY RIVER
35	42	RENFREW	35	43	RUSSELL
35	44	SIMCOE	35	45	STORMONT
35	46	SUDBURY	35	47	THUNDER BAY
35	48	TIMISKAMING	35	49	TORONTO
35	50	VICTORIA	35	51	WATERLOO
35	52	WELLINGTON	35	53	WENTWORTH
35	54	YORK			

Walkthrough: Class of worker (cont'd)

But notice a twist for 1961 data - 4 records per EA!

	PROV	DISTRICT	EA	COUNTY	METUA	RESID	SEX	ALLIND	IND
16	ONTARIO	9	23	8	OSHAWA	FARM	FEMALE	1	
17	ONTARIO	9	23	8	OSHAWA	FARM	MALE	5	
18	ONTARIO	9	23	8	OSHAWA	NON-FARM	FEMALE	60	
19	ONTARIO	9	23	8	OSHAWA	NON-FARM	MALE	236	
20	ONTARIO	9	24	8	OSHAWA	FARM	FEMALE	1	

Walkthrough: Income by community

Average income and income quintiles by constituent community of the amalgamated City of Timmins

Walkthrough: Income by community (cont'd)

Challenge 1: how can we identify the constituent communities in the amalgamated City of Timmins?

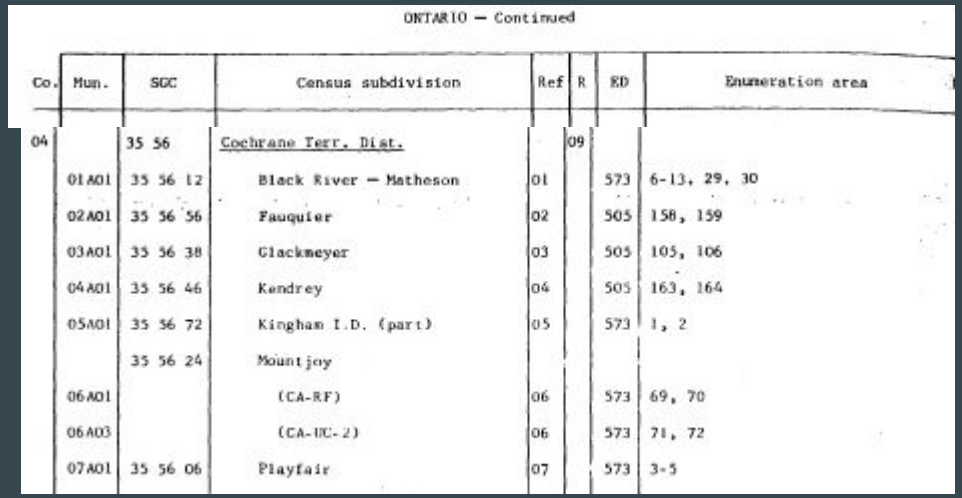
- By using the Official Lists to identify the EAs that comprise these communities

Challenge 2: how can we roll EA-level data up to aggregated communities

- We can sum EA-level incomes quantiles to get community data
- The big problem: can we determine community-level averages from the EA-level averages?

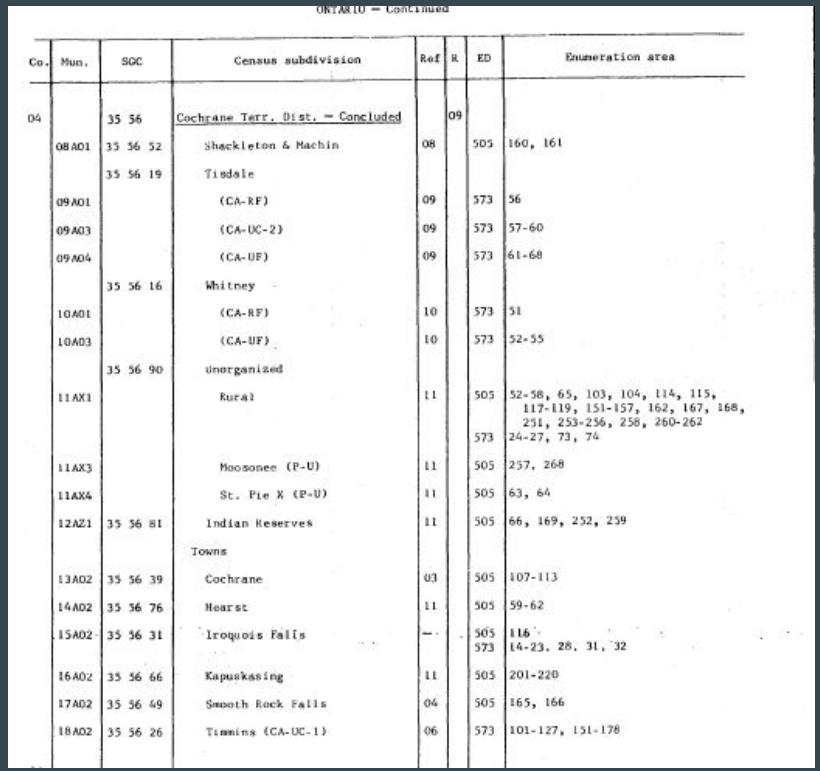
Walkthrough: Income by community (cont'd)

From the Official List (available from DLI EFT):



ONTARIO — Continued

Co.	Mun.	SGC	Census subdivision	Ref	R	ED	Enumeration area
04		35 56	Cochrane Terr. Dist.		09		
	01A01	35 56 12	Black River — Matheson	01		573	6-13, 29, 30
	02A01	35 56 56	Fauquier	02		505	158, 159
	03A01	35 56 38	Glackmeyer	03		505	105, 106
	04A01	35 56 46	Kendrey	04		505	163, 164
	05A01	35 56 72	Kingham I.D. (part)	05		573	1, 2
		35 56 24	Mountjoy				
	06A01		(CA-RF)	06		573	69, 70
	06A03		(CA-UC-2)	06		573	71, 72
	07A01	35 56 06	Playfair	07		573	3-5



ONTARIO — Continued

Co.	Mun.	SGC	Census subdivision	Ref	R	ED	Enumeration area
04		35 56	Cochrane Terr. Dist. — Concluded		09		
	08A01	35 56 52	Shackleton & Machin	08		505	160, 161
		35 56 19	Tisdale				
	09A01		(CA-RF)	09		573	56
	09A03		(CA-UC-2)	09		573	57-60
	09A04		(CA-UF)	09		573	61-68
		35 56 16	Whitney				
	10A01		(CA-RF)	10		573	51
	10A03		(CA-UF)	10		573	52-55
		35 56 90	Unorganized				
	11AX1		Rural	11		505	52-58, 65, 103, 104, 114, 115, 117-119, 151-157, 162, 167, 168, 251, 253-256, 258, 260-262
						573	24-27, 73, 74
	11AX3		Moosonee (P-U)	11		505	257, 268
	11AX4		St. Pie X (P-U)	11		505	63, 64
	12AZ1	35 56 81	Indian Reserves	11		505	66, 169, 252, 259
			Towns				
	13A02	35 56 39	Cochrane	03		505	107-113
	14A02	35 56 76	Hearst	11		505	59-62
	15A02	35 56 31	Iroquois Falls	—		505	116
						573	14-23, 28, 31, 32
	16A02	35 56 66	Kapuskasing	11		505	201-220
	17A02	35 56 49	Smooth Rock Falls	04		505	165, 166
	18A02	35 56 26	Timmins (CA-UC-1)	06		573	101-127, 151-178

Walkthrough: Income by community (cont'd)

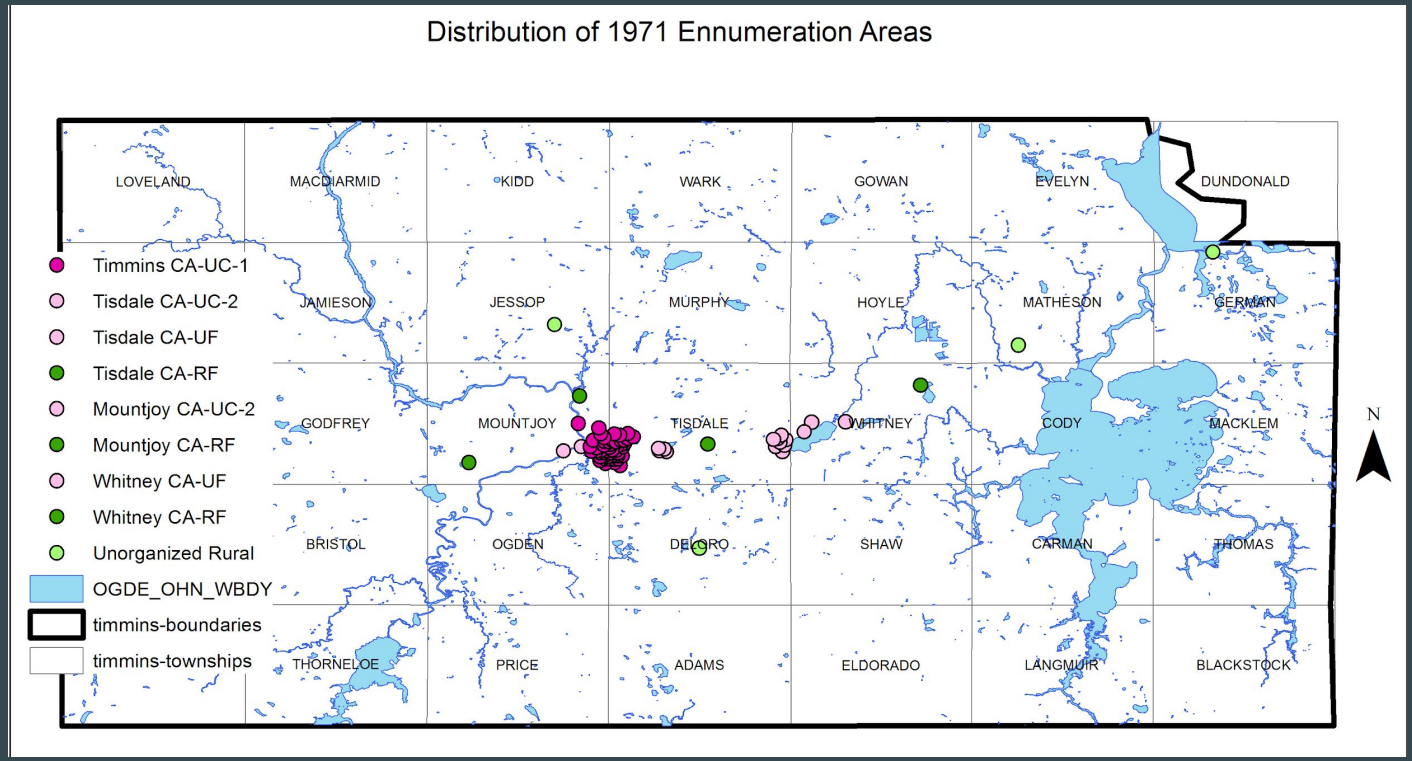
Let's look at that more closely...

ONTARIO — Continued

Co.	Mun.	SGC	Census subdivision	Ref	R	ED	Enumeration area
	17A02	35 56 49	Smooth Rock Falls	04		505	165, 166
	18A02	35 56 26	Timmins (CA-UC-1)	06		573	101-127, 151-178

- Earlier BSTs: many (all?) use *Electoral Districts*, not CDs, to disambiguate EAs
- Standard Geographical Classification (SGC) implemented in 1970 but doesn't appear to structure data the same way as it does now

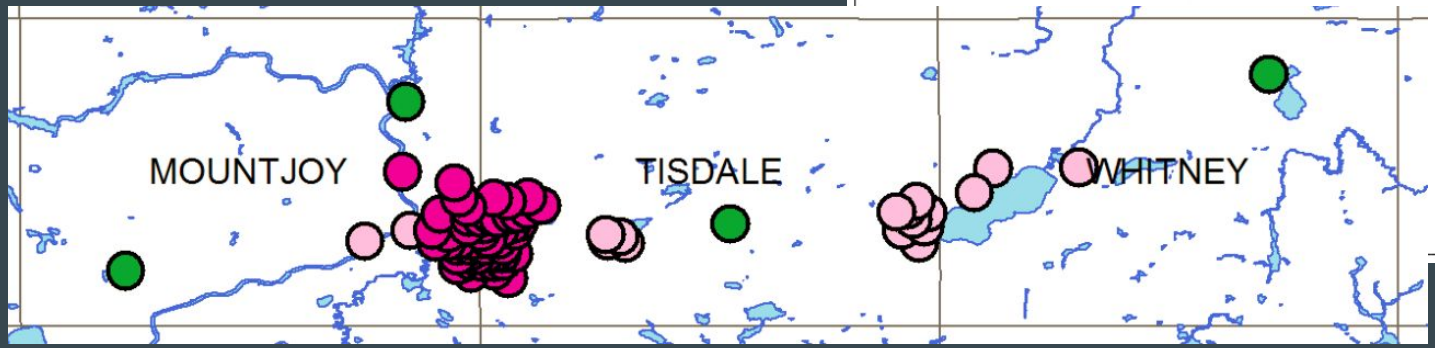
Walkthrough: Income by community (cont'd)

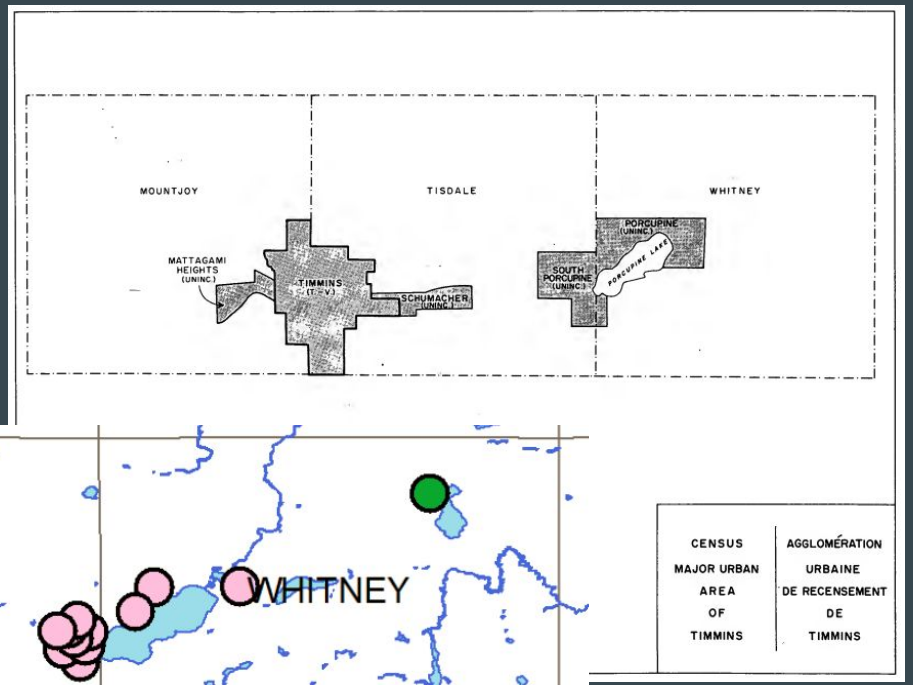


We can use the Geographic Attribute File (GAF) to map the location of the EAs and confirm our assumptions.

Walkthrough: Income by community (cont'd)

That lines up pretty well with the constituent parts of the Timmins CA from a [1961 reference map]





Walkthrough: Figuring out geography (cont'd)

We can create lookup tables for the aggregated communities and use VLOOKUP in Excel to give each EA's record a "locality" value to sum total numbers of individuals per income quantile per community

E	F
ea	Township
26	Unorganized Rural
27	Unorganized Rural
51	Whitney CA-RF
52	Whitney CA-UF
53	Whitney CA-UF
55	Whitney CA-UF
56	Tisdale CA-RF
57	Tisdale CA-UC-2
58	Tisdale CA-UC-2
59	Tisdale CA-UC-2
60	Tisdale CA-UC-2
61	Tisdale CA-UF
62	Tisdale CA-UF
63	Tisdale CA-UF
64	Tisdale CA-UF
65	Tisdale CA-UF
66	Tisdale CA-UF

J78 {=VLOOKUP(I78,locality!$E$2:$F$80,2,FALSE)}

	I	J	M	AB	AC	AD	AE
1	EA	Locality	ANAME	V15	V16	V17	V18
2			Tape variables ->	Tot. Inc. o	Tot. Inc. o	Tot. Inc. o	Tot. Inc. o
3			SPSS syntax variables ->	Males an	Males an	Males an	Males an
4	Enum	Locality	Area name	Zero to	Total i	Total i	Total i
7	121	Timmins CA-UC-1	MR3P5ED73EA121	50	80	50	30
8	122	Timmins CA-UC-1	MR3P5ED73EA122	140	100	65	65
9	123	Timmins CA-UC-1	MR3P5ED73EA123	75	100	50	50

Walkthrough: Income by community (cont'd)

Average income per community:

$$A_C = I_C / P_C$$

Where A_C is average income at the aggregated community level, I_C is the *total* income for the community, and P_C is the total population for the community.

Walkthrough: Income by community (cont'd)

Average income per community:

$$A_C = I_C / P_C > \mathbf{A_C = \Sigma(A_E \times P_E) / \Sigma(P_E)}$$

Where A_E is average income at the EA level, and P_E is the total population for the EA, and Σ is the sum of these for all EAs in a given municipality.

Walkthrough: Income by community (cont'd)

Average income per community:

$$A_C = I_C / P_C$$

BUT!!!

We can try to validate the calculation and there appears to be a problem with some of the data: due to rounding? Due to suppression?

Walkthrough: Income by community (cont'd)

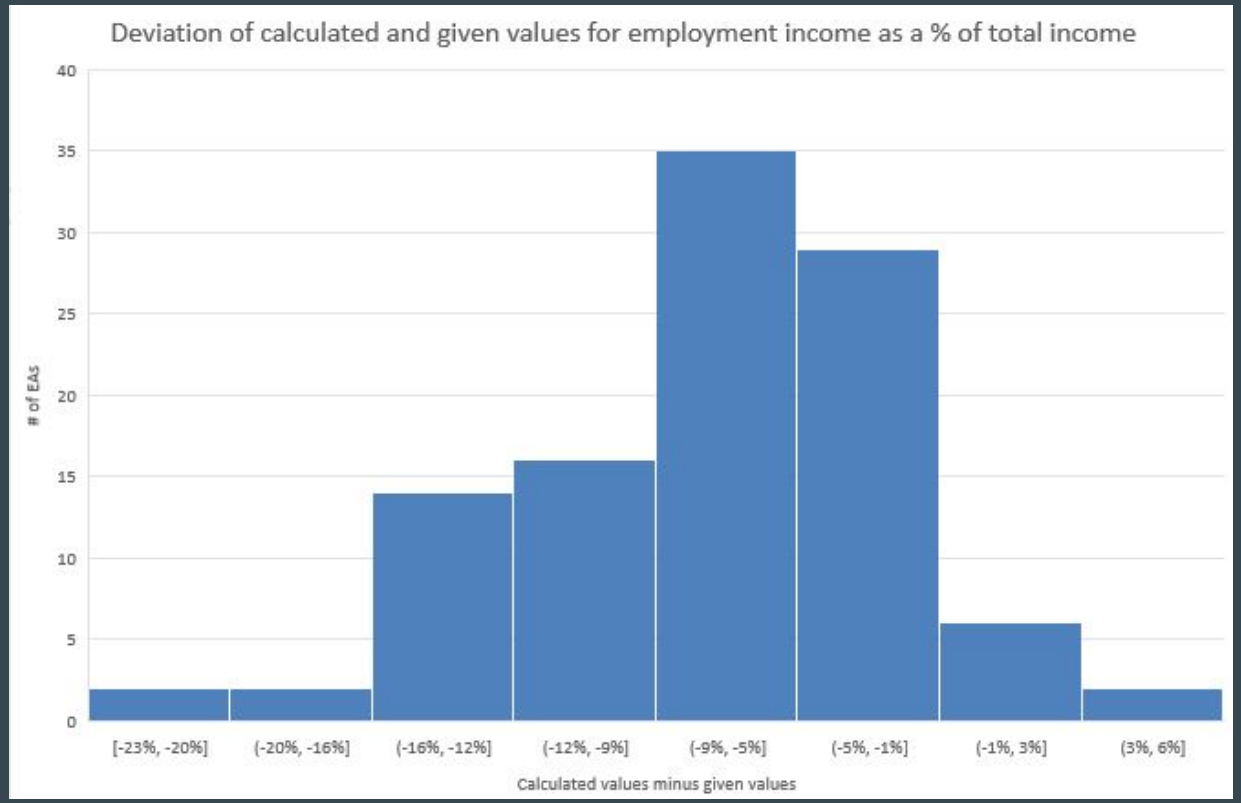
Variables in A2ECN001:

- Number of people earning income (P_I)
- Average income earned (A_I)
- Number of people earning employment income (P_E)
- Average employment income earned (A_E)
- Employment income as a percentage of total income (R)

For each EA, we would expect that $R = (A_E \times P_E) / (A_I \times P_I)$

... but it *isn't*.

Walkthrough: Income by community (cont'd)



Solution? StatCan offers custom tabulations for Census data as early as 1971.

Additional reading - a *very* partial list

- Vince Gray and Liz Hill's "The Academic Data Librarian Profession in Canada: History and Future Directions" (also published in *Databrarianship*)
  - Discusses tape-buying Census consortia which led to the development of the DLI
- Gail Curry's "SPSS Syntax Files: A Do-It-Yourself Primer" and Vince Hill's "Basics of writing SPSS syntax files" (DLI Training presentations)
  - Good intros to SPSS for the uninitiated. Gail's addresses importing fixed-width alphanumeric data without syntax files
- Julie Marcoux's "Basics of Extracting Data with SPSS - Fun with Historical Census" (DLI Training presentation)
  - Discusses PUMFs, which were not addressed in this presentation

What have we (hopefully) learned?

- What we can get from BSTs that we can't get from print sources
- How to parse geographic codes in BSTs
- Variability in data availability and documentation from Census to Census
- How to roll up EA-level data into higher levels of geography - and potential pitfalls in doing so

And,

- Had an opportunity to reflect on the challenges of working with 50+ year old data and documentation, and how this relates to current data preservation initiatives